

**Amendment and Response**

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

---

**IN THE CLAIMS**

Please add claims 31-35.

Please cancel claims 13 and 14.

Please amend claims 1, 11-12, 18, and 21 as follows:

---

- Sub  
B7  
A'*
1. (Currently Amended) A computer-implemented method for mining a document containing dirty text comprising:  
removing an instance of dirty text within said document to produce a cleaned document having a content; and  
performing a data mining operation on said cleaned document [thereby deriving relevant information from said cleaned document and providing a summary of the content of said document. ]
  2. (Original) The method for mining a document containing dirty text as recited in Claim 1, wherein said removing further comprises replacing an instance of dirty text with a standard term.
  3. (Original) The method for mining a document containing dirty text as recited in Claim 1, wherein said removing further comprises removing an instance of computer code from said document.
  4. (Original) The method for mining a document containing dirty text as recited in Claim 1, wherein said removing further comprises removing a table from said document.
  5. (Original) The method for mining a document containing dirty text as recited in Claim 1, wherein said performing a data mining operation further comprises identifying a sentence within said cleaned document by identifying a beginning and an end of said sentence.

**Amendment and Response**

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

---

6. (Original) The method for mining a document containing dirty text as recited in Claim 5, wherein said performing a data mining operation further comprises scoring and ranking said sentence.
7. (Original) The method for mining a document containing dirty text as recited in Claim 6, wherein scoring said sentence further comprises:
- selecting scoring techniques operable for summarizing non-narrative, grammatically incorrect text;
  - selecting scoring techniques operable for summarizing narrative, grammatically correct text; and
  - using said scoring techniques to score said sentence.
8. (Original) The method for mining a document containing dirty text as recited in Claim 7, wherein said method further comprises generating a summary derived from said scored and ranked sentences.
9. (Original) The method for mining a document containing dirty text as recited in Claim 1, wherein said method further comprises selecting a text mining component based upon said data mining operation to be performed.
10. (Original) The method for mining a document containing dirty text as recited in Claim 1, wherein said method further comprises customizing said method by adjusting a parameter value.
11. (Currently Amended) A computer system comprising:
- a bus;
  - a memory unit coupled to said bus; and
  - a processor coupled to said bus, said processor for executing a method for mining a document containing dirty text comprising:
- producing a cleaned document having a content comprising performing a general cleaning of said document by removing an instance of dirty text within said document

**Amendment and Response**

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

---

including instances of misspelling and grammatical errors, and performing a domain and task specific cleaning of said document including removing instances of computer code and tables to produce a cleaned document; and

performing a data mining operation on said cleaned document including providing a summary of the content of said document.

12. (Currently Amended) The computer system as recited in Claim 11, wherein said removing further comprises replacing an instance of dirty text with a standard term.

13. (Cancelled)

14. (Cancelled)

15. (Original) The computer system as recited in Claim 11, wherein said performing a data mining operation further comprises identifying a sentence within said cleaned document by identifying a beginning and an end of said sentence.

16. (Original) The computer system as recited in Claim 15, wherein said performing a data mining operation further comprises scoring and ranking said sentence.

17. (Original) The computer system as recited in Claim 16, wherein scoring said sentence further comprises:

selecting scoring techniques operable for summarizing non-narrative, grammatically incorrect text;

selecting scoring techniques operable for summarizing narrative, grammatically correct text; and

using said scoring techniques to score said sentence.

18. (Currently Amended) The computer system as recited in Claim 17, wherein said method further comprises generating the summary derived from said scored and ranked sentences.

**Amendment and Response**

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

---

- 91
19. (Original) The computer system as recited in Claim 11, wherein said method further comprises selecting a text mining component based upon said data mining operation to be performed.
20. (Original) The computer system as recited in Claim 11, wherein said method further comprises customizing said method by adjusting a parameter value.
21. (Currently Amended) A computer-useable medium having computer-readable program code embodied therein for causing a computer system to perform the steps of:  
removing an instance of dirty text within said document to produce a cleaned document having a content; and  
performing a data mining operation on said cleaned document to provide a summary of said content.
22. (Original) The computer-useable medium of Claim 21, wherein said removing further comprises replacing an instance of dirty text with a standard term.
23. (Original) The computer-useable medium recited in Claim 21, wherein said removing further comprises removing an instance of computer code from said document.
24. (Original) The computer-useable medium recited in Claim 21, wherein said removing further comprises removing a table from said document.
25. (Original) The computer-useable medium recited in Claim 21, wherein said performing a data mining operation further comprises identifying a sentence within said cleaned document by identifying a beginning and an end of said sentence.
26. (Original) The computer-useable medium recited in Claim 25, wherein said performing a data mining operation further comprises scoring and ranking said sentence.

**Amendment and Response**

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

---

27. (Original) The computer-useable medium recited in Claim 26, wherein scoring said sentence further comprises:

selecting scoring techniques operable for summarizing non-narrative, grammatically incorrect text;

selecting scoring techniques operable for summarizing narrative, grammatically correct text; and

using said scoring techniques to score said sentence.

28. (Original) The computer-useable medium recited in Claim 27, wherein said method further comprises generating a summary derived from said scored and ranked sentences.

29. (Original) The computer-useable medium as recited in Claim 21, wherein said method further comprises selecting a text mining component based upon said data mining operation to be performed.

30. (Original) The computer-useable medium as recited in Claim 21, wherein said method further comprises customizing said method by adjusting a parameter value.

31. (New) A computer-implemented method for mining a document containing dirty text comprising:

producing a cleaned document having a content comprising performing a general cleaning of said document by removing one or more instance of dirty text within said document including instances of misspelling and grammatical errors, and performing a domain and task specific cleaning of said document including removing instances of computer code and tables; and

performing a data mining operation on said cleaned document, including determining a sentence score for each sentence of said cleaned document and ranking the sentences from highest to lowest based on the sentence score;

generating a summary of the content of the document using the highest ranked sentences.

**Amendment and Response**

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

---

32. (New) The method of claim 31, wherein determining a sentence score for each sentence includes applying a keyword technique to each sentence.

a' 33. (New) The method of claim 32, wherein determining a sentence score further comprises applying a location technique to each sentence.

34. (New) The method of claim 32, wherein determining a sentence score further comprises applying a semantic similarity technique to each sentence.

35. (New) The method of claim 34, wherein the semantic similarity technique comprises:  
generating a vector associated with each sentence; and  
comparing each vector to every other vector, including defining a cosine of an angle between two vectors and using the cosine of the angle between two vectors to determine whether sentences represented by the two vectors are semantically related.

---